JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 11, 2012
© Mary Ann Liebert, Inc.
Pp. 1251–1264
DOI: 10.1089/cmb.2012.0187

Locally Learning Biomedical Data Using Diffusion Frames

M. EHLER,[1,2] F. FILBIR,[1] and H.N. MHASKAR[3,4]

ABSTRACT

Diffusion geometry techniques are useful to classify patterns and visualize high-dimensional datasets. Building upon ideas from diffusion geometry, we outline our mathematical foundations for learning a function on high-dimension biomedical data in a local fashion from training data. Our approach is based on a localized summation kernel, and we verify the computational performance by means of exact approximation rates. After these theoretical results, we apply our scheme to learn early disease stages in standard and new biomedical datasets.

Key words: graphs and networks, machine learning.

1. INTRODUCTION

As personalized medicine expands, increasingly detailed biomedical data must be integrated to better understand normal function and evolution of multifactorial chronic disease in clinical trials and individual treatment decisions. The complexity of molecular, cellular, and tissue interactions, however, is a fundamental barrier to extracting the complicated relationships that underlie human physiology.

A recent idea, originating in computational harmonic analysis, is to let the data speak for itself. In this approach, one deals typically with high-dimensional, unstructured data. In theoretical analysis, one assumes that the data represents a sample from some *unknown* low-dimensional manifold embedded in a high-dimensional ambient Euclidean space. The objective is then to understand the geometry of this manifold. Thus, statistical techniques have been devised to estimate the dimension of this manifold (Costa and Hero, 2004). A simulation of Brownian motion is expected to reveal the relative neighborhoods of different data points, as well as provide local coordinate systems for the manifold (Jones et al., 2010; Lafon, 2004). See the special issue (Chui and Donoho, 2006) for an introduction to these ideas. Related analysis of graphs is discussed in Pesenson and Pesenson (2010).

In many practical applications, one needs to go beyond an understanding of the manifold and answer queries based on the data. These queries can be modeled mathematically as functions in the (unknown) manifold. This function may be known to us on few training points, and we aim to accurately predict the value of the function of items that are not yet observed. Models for this have been developed as eigenmaps/diffusion maps (Coifman et al., 2005a), multiscale approaches (Coifman et al., 2005b; Gavish et al., 2010;

[1] Helmholtz Zentrum München, Institute of Biomathematics and Biometry, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.
[2] Section on Medical Biophysics, NICHD, National Institutes of Health, Bethesda, Maryland.
[3] Department of Mathematics, California Institute of Technology, Pasadena, California.
[4] Department of Mathematics, Claremont Graduate University, Claremont, California.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Locally Learning Biomedical Data Using Diffusion Frames**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **California Institute of Technology,Department of Mathematics,Pasadena,CA,91125**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Diffusion geometry techniques are useful to classify patterns and visualize high-dimensional datasets. Building upon ideas from diffusion geometry, we outline our mathematical foundations for learning a function on high-dimension biomedical data in a local fashion from training data. Our approach is based on a localized summation kernel, and we verify the computational performance by means of exact approximation rates. After these theoretical results, we apply our scheme to learn early disease stages in standard and new biomedical datasets.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT **Same as Report (SAR)**	18. NUMBER OF PAGES **14**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

Nadler and Galun, 2007), and nonlinear dimension reduction (Belkin and Niyogi, 2003; Chui and Wang, 2010; Roweis and Saul, 2000; Saito, 2008; Tenenbaum et al., 2000; Wu et al., 2010).

Necessarily, any such model cannot be expected to yield perfect reproduction of the actual target function. The subject of approximation theory deals with the intrinsic errors inherent in constructing different kinds of models for the target function. In traditional scenarios, the accuracy of approximation is closely related to the smoothness of the function. Because of this history, many experts in approximation theory nowadays consider the accuracy of approximation itself to be a measurement of smoothness. This viewpoint is particularly useful in our setting, where the manifold is unknown and, therefore, it is impossible to define the smoothness in a classical manner.

The last named author and his collaborators have developed approximation theory tools applicable in the current context in a series of papers (Filbir and Mhaskar, 2010; Filbir and Mhaskar, 2011; Maggioni and Mhaskar, 2008; Mhaskar, 2010; Mhaskar, 2011). A particularly interesting aspect of this theory is a definition of pointwise smoothness of the target function. The research has also enabled us to devise specific algorithms, extending the theory developed for the understanding of geometry, with the property that the rate of convergence of these algorithms in neighborhoods of different points completely characterize local smoothness properties of the target function at those points.

Such local smoothness ideas are particularly useful in biomedicine, as disease progression underlies natural variations, medication leads to abrupt changes in disease progression, and environmental factors vary quickly, so that the query function might not be globally smooth. While late disease stages underlie large variations, the transition from healthy to early pathology can be smooth, leading to query functions that are locally smooth within such early disease transitions.

In this article, we shall review and further develop the salient features of diffusion geometry and approximation theory needed to ''learn locally'' from the acquired data. In contrast to common clustering methods used in biomedicine, we explicitly use that the clusters represent disease stages, i.e., are ordered quantitatively in a progressing fashion. Thus, some unspecific ordering is a-priori fed into the clustering process and is specified in the final result. Moreover, instead of interpolating on the training data, which usually leads to instabilities with large training sets, we allow our algorithm to correct misclassified training points.

The outline is as follows. In Section 2, we briefly discuss two approaches to reconstructing the query function from training data. In Section 3, we present our local learning approach. The numerical implementation is discussed in Section 4. In Section 5, we outline our scheme for the special case in which the manifold is the sphere. We apply our methods to analyze several biomedical datasets in Section 6.

2. APPROXIMATION ON UNKNOWN MANIFOLDS

Let $\mathbb{X}$ be the underlying but unknown compact manifold, endowed with a probability measure μ. The training data $\mathcal{C}=\{y_i\}_{i=1}^M \subset \mathbb{X}$ yield pairs of the form $\{(y_j, z_j)\}_{j=1}^M$, with $z_j \approx f(y_j)$ for the as yet unknown function f. This defines f only on $\mathcal{C}$. The objective is to extend this function to the entire manifold $\mathbb{X}$, including the data already observed and the data not yet available. In deterministic analysis, we do not deal with the noise explicitly, although some probabilistic estimates can be given in addition to deterministic guarantees that account for the noise (Gia and Mhaskar, 2008a).

There are two common approaches to solving this problem. In the first approach, one finds the extension $f_{\mathcal{C}}$ as a solution of some minimization/regularization problem, for example,

$$f_{\mathcal{C}} = \arg \min_{g\in\mathcal{F}}\{||\mathbf{g}-\mathbf{z}|| + \delta||g||_{\mathcal{F}}\}, \tag{1}$$

where δ is a balancing parameter, $\mathcal{F}$ is a suitable class of functions to choose the modeling function g from, $\mathbf{g}$ denotes the vector $(g(y_1), \dots, g(y_M))$, $\mathbf{z}$ denotes the vector $(z_1, \dots, z_M)$, $||\cdot||$ is some norm on, $\mathbb{R}^M$, and $||\cdot||_{\mathcal{F}}$ is a *penalty functional*, commonly the norm on $\mathcal{F}$. We will call Equation (1) the optimization approach.

The other approach is to imagine *a priori* that there is an underlying unknown function f on $\mathbb{X}$, so that $f(y_i)=z_i$, $i=1, \dots, M$. We then seek a model $P \in \mathcal{F}$ so that for some *performance guarantee* $\epsilon_{\mathcal{C}}$,

$$||f-P||_\infty \leq \epsilon_{\mathcal{C}}||f||_{\mathcal{F}}, \tag{2}$$

where $||\cdot||_\infty$denotes the supremum norm. We will call Equation (2) the approximation approach.

The optimization approach has the advantage that the penalty functional may be chosen to reflect some domain-specific knowledge about the target function. Also, even if one does not expect any function underlying the phenomenon, one gets some smooth model to work with. On the other hand, when we do not know with absolute certainty any physical model that underlies the data, and are seeking a function on $\mathbb{X}$ as a model anyway, then it is more natural to assume that there is an underlying function from which the data is sampled, even though the function itself is yet unknown, so that the approximation approach seems more natural.

We noted some comparisons between the two approaches. First, we the optimization approach does not necessarily imply any performance guarantees of Equation (2). Moreover, the value of the regularization functional depends upon the data. There are no bounds to how large this value might get as more and more data are introduced. Finally, there are common computational issues such as local minima, convergence of the algorithms, and convergence of the minimizers $f_C$as the data becomes dense on the manifold. All of these issues are completely avoided in the approximation approach.

We will demonstrate below that in the approximation approach, we can construct a linear operator with mathematical performance guarantees of Equation (2). We do not need to solve any minimization problem, so that all the computational issues mentioned above are avoided altogether. Moreover, we can design this operator in a manner that its performance guarantees are automatically better on regions of $\mathbb{X}$ where the target function is ‘‘smoother.’’ This does not involve a careful detection of edges and partitioning of $\mathbb{X}$.

3. LOCAL APPROXIMATION OF THE QUERY FUNCTION

3.1. Local smoothness classes

Before we can define local smoothness of a function $f : \mathbb{X} \to \mathrm{R}$, we must specify a few more objects. Let $\{\varphi_k\}_{k=0}^{\infty}$ be the eigenfunctions of the Laplace-Beltrami operator Δ on $\mathbb{X}$, and $\{\ell_k^2\}_{k=0}^{\infty}$ the associated eigenvalues, ordered in a nondecreasing way with $\ell_0 = 0$ and $\ell_k \to \infty$ as $k \to \infty$. The space of *diffusion polynomials* up to degree N is $\Pi_N := \mathrm{span}\{\varphi_k : \ell_k < N\}$. Further technical details are discussed in Appendix A. Note that $\{\varphi_k\}_{k=0}^{\infty}$ was replaced by a more general orthonormal basis for $L_2(\mathbb{X},\mu)$ in Filbir and Mhaskar (2010, 2011).

The object of interest in approximation theory is the *degree of approximation* of the target function:

$$E_N(f) = \min_{P \in \Pi_N} \|f - P\|_\infty. \tag{3}$$

Equation (3) measures the best error achievable if one wishes to use Π_N as the model for f, and wishes the error to be small at each point of $\mathbb{X}$. It turns out that the rate at which the quantity $E_N(f)$ decreases to 0 as $N \to \infty$ is closely related to the smoothness of f. Thus, if f is smooth enough so that $\Delta^r f \in \mathscr{C}(\mathbb{X})$ for some integer $r \geq 1$, then

$$E_N(f) \leq \frac{c}{N^{2r}} \|\Delta^r f\|_\infty,$$

where $c > 0$ is a constant. Here, $\mathscr{C}(\mathbb{X})$ denotes the continuous functions on $\mathbb{X}$ endowed with the supremum norm. In practice, it is quite common to find an approximation of f by ad hoc means. This gives rise to the question that if one finds $E_N(f) \downarrow 0$ at a certain rate, is it because f inherits a certain smoothness? Thus, the correct notion of smoothness required to answer this question is defined in terms of a regularization functional (known in some branches of analysis as the K-functional). If $r \geq 1$ is an integer, this functional is defined by

$$K_r(f, \delta) = \inf\{ \|f - g\|_\infty + \delta^{2r} \|\Delta^r g\|_\infty\}, \qquad \delta > 0,$$

where the infimum is taken over all g for which $\|\Delta^r g\|_\infty < \infty$. In the terminology introduced before, this measures the approximation of f by smooth functions g while controlling the growth of $\Delta^r g$ with the help of the regularization parameter δ. It is known (Maggioni and Mhaskar, 2008) that if $s > 0$ and $r > s/2$ is any integer, then

$$E_N(f) = \mathcal{O}(N^{-s}) \text{ as } N \to \infty \text{ if and only if } K_r(f, \delta) = \mathcal{O}(\delta^s) \text{ as } \delta \to 0.$$

Standard approximation theory arguments show that if $\gamma < s/2$ is an integer, and $\beta = s - 2\gamma$, then

$$E_N(f) = \mathcal{O}(N^{-s}) \text{ if and only if } \Delta^\gamma f \in \mathscr{C}(\mathbb{X}) \text{ and } K_r(\Delta^\gamma f, \delta) = \mathcal{O}(\delta^\beta).$$

In particular, the choice of $r > s/2$ is not critical except that the constants involved in the $\mathcal{O}$ relations will depend upon the choice of r. This leads us to define the (global) smoothness class W^s directly in terms of the quantities $E_N(f)$ as

$$W^s(\mathbb{X}) := \{f \in \mathscr{C}(\mathbb{X}) : E_N(f) = \mathcal{O}(N^{-s})\},$$

endowed with the norm $\|\cdot\|_{W^s} := \|\cdot\|_\infty + \|(N^s E_N(f))_N\|_\infty$. We will think of s as the parameter measuring the smoothness of the function f.

In the context of data-defined manifolds, we do not explicitly know any formulas for the local coordinate charts. Therefore, it is not easy to define the notion of derivatives. Nevertheless, we can redefine the notion of an infinitely differentiable function on our unknown manifold as membership in *every* smoothness class $W^s(\mathbb{X})$. Thus, the set of all infinitely differentiable functions is denoted by $\mathscr{C}^\infty(\mathbb{X}) = \bigcap_{s>0} W^s(\mathbb{X})$. If $x_0 \in \mathbb{X}$, we will define the local smoothness of f at x_0 by the natural windowing construction. Thus, we say that $f \in W^s_{x_0}(\mathbb{X})$if there exists a neighborhood U of x_0 such that for every $\phi \in \mathscr{C}^\infty(\mathbb{X})$, supported on U, $\phi f \in W^s(\mathbb{X})$.

3.2. *Localized summation kernels*

We can clearly expand any $f \in L_2(\mathbb{X}, \mu)$ in the orthonormal basis $\{\varphi_k\}_{k=0}^\infty$ by $f = \sum_{k=0}^\infty \langle f, \varphi_k\rangle \varphi_k$. To motivate our scheme presented later, we manipulate this expression without caring about convergence and interchanging limits, so that we derive

$$f(x) = \sum_{k=0}^\infty \int_{\mathbb{X}} f(y)\varphi_k(y) d\mu(y) \varphi_k(x) = \int_{\mathbb{X}} f(y)\Phi(x, y) d\mu(y), \tag{4}$$

where we formally use $\Phi(x, y) = \sum_{k=0}^\infty \varphi_k(y)\varphi_k(x)$. This representation requires knowledge of f on the entire manifold $\mathbb{X}$. To reconstruct f from the data only, we must replace the integral with a finite sum over data points and localize the kernel Φ so that $f(x)$ is determined by its values in a neighborhood around x. In this section, we shall make these ideas mathematically precise and construct a linear operator based on the data to derive $P \in \Pi_N$ that essentially minimizes Equation (3).

First, we define

$$\Phi_N(x, y) := \sum_{k=0}^\infty h\left(\frac{\ell_k}{N}\right)\varphi_k(x)\varphi_k(y), \quad \text{for all } x, y \in \mathbb{X}, \tag{5}$$

where $h : \mathbb{R} \to [0, 1]$ satisfies $h(t) = 1$ if $|t| \leq 1/2$ and $h(t) = 0$ if $|t| \geq 1$, see Figure 1 for a typical function h and Appendix A for a few more technical conditions. Due to the ‘‘cut-off’’ function h, Equation (5) is a

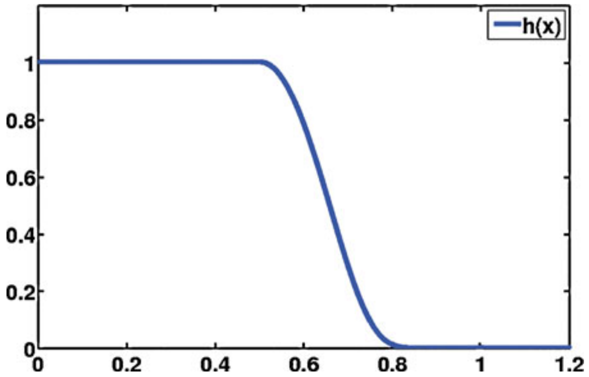

FIG. 1. Filter function: A typical filter function h that can be used in Equation (5).

finite sum, and Φ_N, as a function of only one variable x or y, is contained in Π_N. Under the technical assumptions collected in Appendix A, the estimate

$$|\Phi_N(x,y)| \lesssim \frac{N^{\alpha}}{\max\,(1,(N\rho(x,y))^{S}}, \tag{6}$$

holds, where α is the dimension of $\mathbb{X}$, ρ the metric on $\mathbb{X}$, and S any integer (Filbir and Mhaskar, 2010; Maggioni and Mhaskar, 2008; Mhaskar, 2011). Here means that there is an absolute positive constant on the right-hand side such that the inequality holds. This localization property (6), very much like multi-scale approaches with wavelets, enables local analysis of functions on the manifold.

If $f \approx P$, then the we also have $f\Phi_N(x,\,\cdot\,) \approx P\Phi_N(x,\,\cdot\,)$. Naturally, we would expect that the product of two polynomials is again a polynomial of a larger degree, so that $P\Phi_N(x,\,\cdot\,) \in \Pi_{aN}$, for some fixed $a > 0$. Here, we only need the weaker product assumptions. (Appendix A), implying that $P\Phi_N(x,\,\cdot\,)$ can be sufficiently well approximated with polynomials in Π_{aN}. To replace the integral in Equation (4) with a finite sum, we need a quadrature formula that is exact for polynomials up to degree aN: If our data $\mathcal{C} = \{y_i\}_{i=1}^{M}$ are sufficiently dense in $\mathbb{X}$ (Appendix B), then there are quadrature weights $\{\omega_i\}_{i=1}^{M}$ such that

$$\int_{\mathbb{X}} P(x)d\mu(x) = \sum_{j=1}^{M} \omega_j P(x_j), \quad \text{for all } P \in \Pi_{aN}.$$

Since $\varphi_0 \equiv 1$, these weights can be computed by solving the linear system of equations

$$\sum_{j=1}^{M} \omega_j \varphi_k(x_j) = \delta_{k,0}, \quad \text{for all } k = 0, 1, 2, \ldots, aN.$$

We refer to Appendix B and Filbir and Mhaskar (2010) for details on the existence of such quadrature weights with respect to the training data.

We now replace the right-hand side of Equation (4) with

$$\sigma_N(f,x) : = \sum_{j=1}^{M} \omega_j f(y_j)\Phi_N(x,y_j). \tag{7}$$

To study the asymptotics for $N \to \infty$, we call for a sequence of training sets $\mathcal{C}_N$ that induce quadrature formulas of strength aN. We shall verify in Appendix C that, for $f \in W^{s}_{y_0}(\mathbb{X})$, there is $\delta > 0$, such that,

$$\sup_{x \in B_\delta(y_0)} |f(x) - \sigma_N(f,x)| \lesssim N^{-s}, \tag{8}$$

where $B_\delta(y_0) \subset \mathbb{X}$denotes a ball of radius δ around y_0. Thus, when f is locally smooth in a neighborhood around y_0, then we can locally reconstruct f from the training data. The analogous result for functions that are globally smooth is contained in Filbir and Mhaskar (2010, 2011), Maggioni and Mhaskar (2008), and Mhaskar (2011), which contains local estimates, but the approximand requires global knowledge of f and is not purely defined through the training data only.

4. SPECIFYING φ_K NUMERICALLY

To design a numerically feasible algorithm, we still need to compute a suitable orthonormal basis $\{\varphi_k\}_{k=0}^{\infty}$ for $L_2(\mathbb{X},\mu)$, while the manifold $\mathbb{X}$ is not explicitly known to us. In a typical situation, we may have little training data and a much larger but finite collection $\{y_i\}_{i=1}^{M}$ of data lying on the manifold. These data shall be used to approximate an orthonormal basis$\{\varphi_k\}_{k=0}^{\infty}$.

The eigenfunctions of the Laplace-Beltrami operator $\Delta_{\mathbb{X}} = \mathrm{div}\nabla$ on the manifold form an orthonormal basis that satisfies the technical assumptions needed in Appendix A. In order to compute them numerically, we build the graph Laplacian from a finite set of points on the manifold as follows: By using data points $\{y_i\}_{i=1}^{M} \subset \mathbb{X}$, the standard heat kernel $k_\varepsilon(x,y) = e^{-\|x-y\|^2/2\varepsilon}$, $\varepsilon > 0$, induces the weight matrix $\mathcal{W}_M^{\varepsilon}$ defined by $\mathcal{W}^{\varepsilon}_{M;i,j} = k_\varepsilon(y_i,y_j)$. We build the diagonal matrix $D^{\varepsilon}_{M;i,i} = \sum_{j=1}^{M} \mathcal{W}^{\varepsilon}_{M;i,j}$ and define the (unnormalized) graph Laplacian as $L^{\varepsilon}_{M} = \mathcal{W}^{\varepsilon}_{M} - D^{\varepsilon}_{M}$. For $M \to \infty$ and $\varepsilon \to 0$, the eigenvalues and interpolations of the eigenvectors of L_M^ϵconverge toward the "interesting" eigenvalues and eigenfunctions of $\Delta_{\mathbb{X}}$ when $\{y_i\}_{i=1}^M$are uniformly distributed on $\mathbb{X}$. See Appendix D for the details and technical assumptions. If $\{y_i\}_{i=1}^M$ are distributed according to the density p, then the graph Laplacian approximates the elliptic Schrödinger-type operator $\Delta + \frac{\Delta p}{p}$ (Coifman et al., 2005b; Nadler et al., 2006), whose eigenfunctions also form an orthonormal basis for $L_2(\mathbb{X}, \mu)$.

5. LOCALIZED KERNELS ON THE SPHERE

The data of one of our applications lie on the sphere S^{d-1}, so that we specify our approach for $\mathbb{X} = S^{d-1}$ (Gia and Mhaskar, 2006; Gia and Mhaskar, 2008b). For $f : S^{d-1} \to \mathrm{R}$, let $\check{f}(x) : = f(x/ \|x\|)$ be its extension to $\mathbb{R}^d \backslash \{0\}$. The Laplace-Beltrami operator $\Delta_{S^{d-1}}$ is $\Delta_{S^{d-1}} f = (\Delta_{\mathrm{R}^d} \check{f})_{|S^{d-1}}$, and the set of spherical harmonics $\mathcal{H}_k^d$ is formed by the homogeneous harmonic polynomials p on $\mathbb{R}^d$of degree k restricted onto the sphere S^{d-1}. In other words, p is a polynomial whose monomials have all the same total degree k and$\Delta_{\mathbb{R}^d} p = 0$. The eigenspaces of $\Delta_{S^{d-1}}$ are $\mathcal{H}_k^d$ with eigenvalues $k(k + d - 2)$, respectively, and the dimension of $\mathcal{H}_k^d$ is $m_k := \binom{k+d-1}{k} - \binom{k+d-3}{k-2}$. If we choose an orthonormal basis $\{\varphi_{k,j}\}_{j=1}^{m_k}$ for $\mathcal{H}_k^d$, then $\sum_{j=1}^{m_k} \varphi_{k,j}(x)\varphi_{k,j}(y)$ does not depend on the choice of $\{\varphi_{k,j}\}_{j=1}^{m_k}$ and coincides with $P_k(\langle x, y\rangle)$, where P_k is the Gegenbauer polynomial of degree k and parameter $d/2 - 1$, sec. Appendix E.1 and Stein and Weiss (1971). Therefore, we can explicitly compute

$$\Phi_N(x, y) = \sum_{k=0}^{N} h\left(\frac{\sqrt{k(k+d-2)}}{N}\right) P_k(\langle x, y\rangle).$$

To derive $\sigma_N(f,x)$ in Equation (7), we still need to determine the quadrature weights $\{\omega_j\}_{j=1}^n$, which we shall properly describe in Appendix E.2. Here, we only present a heuristic approach. Since $\{P_k(\langle x, \cdot\rangle)\}_{x \in S^{d-1}}$ generates $\mathcal{H}_k^d$, we aim to solve, for R as large as possible,

$$\sum_{j=1}^{n} \omega_j P_k(\langle x, x_j\rangle) = \delta_{k,0}, \qquad \text{for all } k = 0, \ldots, R,$$

and $m_k = \dim(\mathcal{H}_k^d)$ random choices of $x \in S^{d-1}$. This leads to a linear system of equations whose solution is reasonably close to exact weights in practice, but only for small parameters d, n, and R. If any of these parameters is not small, then the problem becomes numerically unstable, and we need to follow the approach presented in Appendix E.

If the manifold is known, as in the case of the sphere, many types of basis functions can be constructed explicitly. For instance, spherical wavelets were considered in Antoine and Vandergheynst (1995), Dahlke et al. (1995, 2004), and Starck et al. (2006), which can capture multi-scale structure. However, replacing the eigenfunctions $\{\varphi_k\}_{k=0}^\infty$ of the Laplace-Beltrami operator would require checking in each case if all the conditions in Appendix A are satisfied. Therefore, we shall not follow this path and, here, exclusively use the eigenfunctions of the Laplace-Beltrami operator.

6. APPLICATIONS

We use the developed approximation scheme to cluster biomedical data into disease stages $\{C_l\}_{l=0}^L$. Therefore, the classes (disease stages) have a natural ordering, and we assign a meaningful number c_l to cluster C_l such that $0 \le c_0 < c_1 < \ldots < c_L$. Hence, $|c_i - c_j|$ represents the distance between C_i and C_j. The query function$f$ is defined on the training data by$f(x_i) = c_l$ if and only if $x_i \in C_l$. The approximand $\sigma_N(f,x)$ induces a nearest neighbor classification on the entire dataset by the proximity of $\sigma_N(f,x)$ to any of the numbers $c_0, \ldots, c_L$.

We first compare our proposed approach to widely used classification methods on two standard biomedical datasets. After this verification, we use our scheme to analyze multispectral retinal images of age-related macular degeneration (AMD) patients. All eye-related data were collected by our collaborators at the National Eye Clinic at the National Institutes of Health (Bethesda, NIH; Maryland).

6.1. Standard biomedical datasets

6.1.1. Cleveland Heart Disease Database: learning disease stages. The Cleveland Heart Disease Database (CHDD) (Detrano et al., 1989) contains 297 patterns with 13 attributes and is grouped into five progressive heart disease stages (values 0,1,2,3,4), where 0 corresponds to normal heart conditions. We removed six patterns due to missing values. Due to homeostasis and its failure in progressed disease, we expect the query function f to be smoother on normal heart conditions and early disease stages, while later stages may form a more heterogeneous group. We compare our method to support vector machines (SVM), in which clusters are derived through sequential binary clustering. Clusters are evaluated by means of binary false-positive or false-negatives for each disease stage. Indeed, our proposed method recovers f consistently better than SVM for the values 0, 1, and 2 when dealing with few training points (Table 1). As expected, our kernel performs poorly on the stages 3 and 4, implying the lack of smoothness of f within these progressed stages. The transition from 0 to 2 seems to be steered by a smoother process, which is reflected by a smoother query function yielding better results than SVM methods.

6.1.2. Wisconsin Breast Cancer Database: data completion. After removing missing values, the Wisconsin Breast Cancer Database (WBCD; original) (Wolberg and Mangasarian, 1990) contains 683 patterns with 9 attributes. We aim to predict quantitative attributes. In fact, we randomly select 200, 300, and 400 training points to learn the attribute ''clump thickness'' (ranging from 1 to 10) and aim to predict its values on the remaining data. We call it a hit when the prediction is within a radius of 1, i.e., if the measured size was 3, the predictions in the interval [2, 4] are counted as a hit. The excellent performance of our proposed method by means of sensitivity (number of hits divided by the cluster size) for few training data is shown in Table 2.

6.2. Age-related macular degeneration

Age-related macular degeneration is the most common cause of blindness among the elderly population in the western world (Chew et al., 2009; Coleman et al., 2008; Krishnadev et al., 2010). Aging of the human retina is universally associated with microscopic changes within the retinal pigment epithelium (RPE), including increased number and volume of fluorescent lipofuscin granules (Meyers et al., 2004). In a

Table 1. Sensitivity Analysis for Cleveland Heart Disease Database

Stage 0	*SVM linear*	*SVM Gaussian*	*Local kernel*
CHDD, 40	73%	71%	79%
CHDD, 100	78%	80%	83%
CHDD, 200	93%	95%	92%
Stage 1			
CHDD, 40	69%	68%	73%
CHDD, 100	73%	75%	80%
CHDD, 200	88%	92%	85%
Stage 2			
CHDD, 40	64%	62%	71%
CHDD, 100	68%	71%	75%
CHDD, 200	86%	85%	81%
Stage 3			
CHDD, 40	61%	60%	54%
CHDD, 100	65%	66%	59%
CHDD, 200	78%	79%	69%
Stage 4			
CHDD, 40	57%	55%	52%
CHDD, 100	63%	59%	54%
CHDD, 200	72%	69%	63%

Sensitivity (one minus false negative rate) for disease stages 0 to 4 using the dataset CHDD in Section 6.1.1. with 40, 100, and 200 training points, averaged over 50 instances for each method. Our local kernel method performs better for few training data than SVM on disease stages 0, 1, and 2, in which we expect the query function to be relatively smooth. CHDD, Cleveland Heart Disease Database.

Table 2. Sensitivity Analysis for the Wisconsin Breast Cancer Database

	SVM linear	*SVM Gaussian*	*Local kernel*
WBCD, 100	74%	72%	80%
WBCD, 200	76%	79%	84%
WBCD, 300	92%	93%	90%

Sensitivity analysis for the data set WBCD in Section 6.1.2, averaged over clump thickness and 50 instances for each method. Our local kernel method yields high sensitivity compared to other methods when there are only few training data. WBCD, Wisconsin Breast Cancer Database.

majority of Americans over the age of 60, the earliest clinical signs of RPE dysfunction are observed in color fundus photographs as *drusen*—bright highly reflective extracellular deposits between the RPE and its basement membrane. Macular drusen increase in number and size with advancing age in epidemiological studies and larger, irregular-shaped, perifoveal drusen (‘‘soft’’) are considered to confer the greatest risk for progression to advanced AMD. Through many years of large-scale studies of the natural history of AMD and controlled prevention trials, clinical observations of fundus photographs suggest that people with soft (larger than 150 microns and irregularly shaped) drusen are at high risk for progressing to advanced AMD. Currently, pathologists in reading centers classify drusen based on size and shape (Bird et al., 1995) in reflection color fundus images. There is demand for automated analysis tools that allow for quantitative prediction of disease progression.

6.2.1. Retinal multi-spectral imaging. The retina is a multilayer neural tissue, uniquely suited for noninvasive optical imaging with high resolution. The first author, together with his collaborators at NIH, developed noninvasive multispectral fluorescence imaging of the human retina by adding selected interference filter sets to standard fundus cameras, allowing the monitoring of early changes within the RPE via the fluorescent lipofuscin granules (Dobrosotskaya et al., 2010, 2011; Ehler et al., 2010, 2011a, 2011b; Kainerstorfer et al., 2010a, 2010b, 2011). If $F(\Lambda,\lambda)$ denotes the measured autofluorescence, where Λ is the excitation and λ the emission wavelength, then the Beer-Lambert law for the double-path yields

$$F(\Lambda,\lambda)=I(\Lambda)\Phi(\Lambda,\lambda)e^{-(D(\Lambda)+D(\lambda))},$$

where D is the integrated absorbance (optical density) of the tissue the light travels through, $\Phi(\Lambda,\lambda)$ is the fluorescence efficiency of lipofuscin, and $I(\Lambda)$ the radiant power of the excitation light. For each of the six patients, four excitation filters with two emission filters and trifold imaging lead to 24 images (400 × 400 pixels) that are aligned by applying the commercial software i2k Align. We de-noise in z-direction by principal component analysis, keeping five eigenvectors capturing more than 98% of the dataset's variance. Spatial noise is reduced by averaging each pixel with its eight neighbors. Since the pathological changes are related to fractional changes of the autofluorescence, we can normalize the 160,000 vectors to lie on the sphere S^4.

The principal component analysis and normalization also helps to compare pixels across patients. The drusen classification scheme presented in van Leeuwen et al. (2003; see also Bird et al., 1995) leads to 8 progressing stages. We partition them into four classes. An expert grader labeled few spatial regions, which led to 2,000 training pixels in three patients that were each labeled with one of the four partitions. Thus, we have 6,000 training vectors total, and pixels in a single patient can have distinct labels. We shall learn locally the pixel classification into drusen classes. Note that our method relies on the multispectral components of drusen in autofluorescence images rather than the spatial shape that is seen in reflection color fundus images.

After the learning step, we mark a region of interest (ROI) in six patients (including the three patients used for learning). To classify pixels from the ROI into the four partitions, we merge the ROI pixels of the patient under consideration with the ROI pixels of the three learned patients to form the data on the manifold. The labeled regions are the training data, and since the pixel vectors lie on the sphere, we can follow the approach in Section 5 to derive $\sigma_N(f,x)$ (Fig. 2). The majority of the pixels within the ROI would define the drusen class of the patient. It should be mentioned that the size of the ROI influences the classification scheme, and we obtained good results with the center in the fovea and extending to 5 degrees, consistent with common image analysis in ophthalmology.

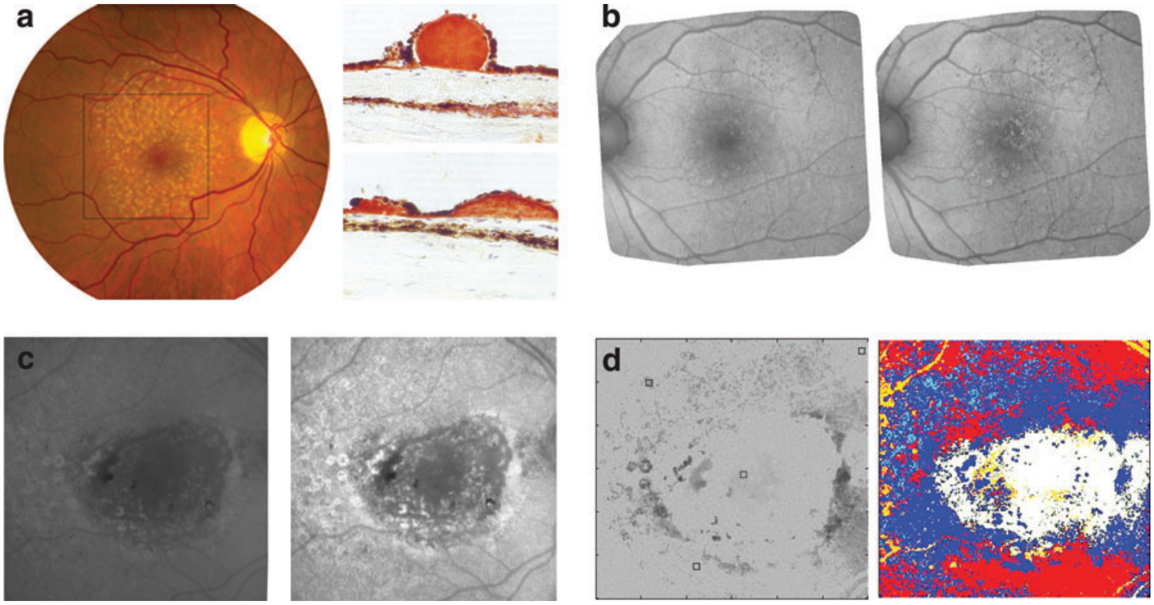

FIG. 2. Drusen classification. The foveal class outnumbers the other classes in (d) and determines the patient's class. (a, left) Drusen in color fundus image; (a, right) Drusen in cross-section of a monkey retina. (b) Two spectral images of pre-advanced AMD patient. (c) Two spectral images of advanced AMD patient. (d, left) Squares show class centers; (d, right) four classes, where dark blue is not assigned to any class.

Next, we explore the influence of the size of the training data. We only use a fraction of the pixels that were originally labeled by our grader. The individual pixels are selected by a random sampling. Figure 3 shows that we require a critical number of training pixels and from there on, we obtain stable classification results.

7. CONCLUSIONS

To facilitate the analysis of complex biomedical data, we developed the mathematical foundations for a numerical algorithm that enables global and local data analysis integratively. After we validated our

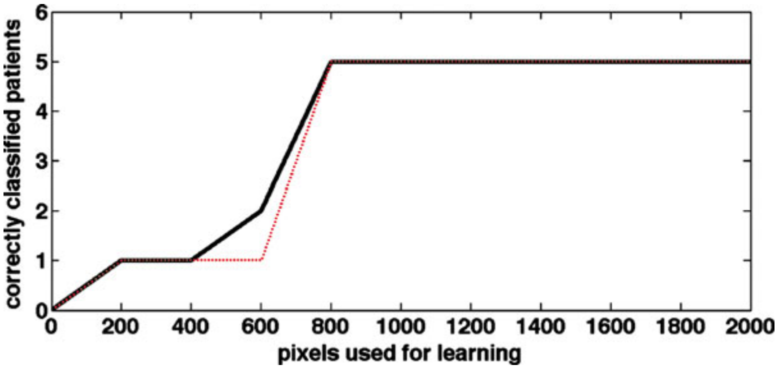

FIG. 3. The increments of the training pixels are in steps of 200. The y-axis counts the correct classifications. Note that we simply removed random pixels from the entire set of 3 × 2000 training pixels, so that each smaller training set is contained in the larger one. We repeated this process over 20 runs to avoid random anomalies, which led to the black curve. The most typical curve in a single run is depicted in red. For few training data, we only classify one single patient correctly. Increasing the size of the training pixels enables consistently correct classifications of all but one patient (consistently the same one). It seems that the critical number of training pixels is between 600 and 800 pixels.

approach on two standard datasets, we aimed to classify and predict disease progression in AMD patients. Drusen were classified in multispectral retinal image sets that enabled quantitative measurements of advanced pathological changes. Clearly, our new mathematical approach to identifying new components controlling AMD progression is preliminary and needs further validation to claim its usefulness in general terms. We anticipate improvements through synergies of multiple analysis schemes with multi modal data so that our proposed analysis could be part of an iterative process.

8. APPENDIX

Appendix A. Technical assumptions

For the sake of readability, we list simple conditions used in Section 3 that can be further weakened (Filbir and Mhaskar, 2010; Maggioni and Mhaskar, 2008). The manifold $\mathbb{X}$ is supposed to be a smooth, compact, and connected Riemannian manifold without boundary. We point out that the following technical conditions are satisfied when $\{\varphi_k\}_{k=0}^{\infty}$ are the eigenfunctions of the Laplace-Beltrami operator on $\mathbb{X}$, $\{\ell_k^2\}_{k=0}^{\infty}$ its eigenvalues, and μ the standard Riemannian probability measure (see Filbir and Mhaskar, 2010, 2011; Maggioni and Mhaskar, 2008; Sikora, 2004 for details).

A.1. Assumption on the volume of balls. If $\alpha > 0$ is the dimension of $\mathbb{X}$, then we assume that $\mu(B_r(x)) \lesssim r^{\alpha}$, for all $x \in \mathbb{X}$ and $r > 0$.

A.2. Assumption on products of polynomials. There is $a \geq 2$ such that, for all $f, g \in \Pi_N$, their product fg is contained in Π_{aN}. In fact, we only need the weaker condition saying that, for

$$\epsilon_N := \sup_{\ell_j,\, \ell_k \leq N} \operatorname{dist}(\varphi_j \varphi_k, \Pi_{aN})_{\infty},$$

we have $N^c \epsilon_N \to 0$ as $N \to \infty$, for every $c > 0$.

A.3. Assumption on the growth of the heat kernel. For

$$K_t(x, y) = \sum_{j=0}^{\infty} \exp(-\ell_j^2 t) \varphi_j(x) \varphi_j(y),$$

let $|\partial^{\gamma} K_t(x, y)| \lesssim t^{-\alpha/2-\gamma} \exp(-c\rho(x, y)^2/t)$,for all $t \in (0, 1]$, $x, y \in \mathbb{X}$, where c is a constant, $\gamma = 0, 1$, and ∂ is a differential operator of first order.

A.4. Assumption on the filter. Let: $\mathbb{R}_{\geq 0} \to \mathbb{R}$ be infinitely often differentiable and a nonincreasing function such that $h(t) = 1$ if $t \leq 1/2$ and $h(t) = 0$ if $t \geq 1$. For instance, we can choose

$$h(x) = \begin{cases} 1, & x \leq 1/2, \\ \exp\left(\frac{(x-\frac{1}{2})^2 (2x^2 - 2x - 1)}{x^2 (x-1)^2}\right), & 1/2 \leq x \leq 1, \\ 0, & 1 \leq x, \end{cases}$$

see Figure 1.

Appendix B. Existence of quadrature weights

For training data $\mathcal{C} = \{x_j\}_{j=1}^{n} \subset \mathbb{X}$, let $\delta_{\mathcal{C}} := \sup_{x \in \mathbb{X}} \operatorname{dist} \rho(x, \mathcal{C})$ and $q_{\mathcal{C}} := \min_{i \neq j} \rho(x_i, x_j)$. If $\mathbb{X}$ is well covered by $\mathcal{C}$, i.e., $\delta_{\mathcal{C}} \leq 2q_{\mathcal{C}} \leq 2\delta_{\mathcal{C}}$, then there exists a constant c with the following property: for $N \leq c/\delta_{\mathcal{C}}$, there are positive numbers $\{\omega_j\}_{j=1}^{n}$, such that the cubature formula $\int_{\mathbb{X}} P(x) d\mu(x) = \sum_{j=1}^{n} \omega_j P(x_j)$ holds, for all $P \in \Pi_N$ (Filbir and Mhaskar, 2010). Note that the weights satisfy $\sum_{j=1}^{n} \omega_j = 1$ and hence are bounded since $\varphi_0 \equiv 1$.

Appendix C. Proof of local approximation

Given $f \in W_{y_0}^{s}(\mathbb{X})$, pick $\delta > 0$ such that $\phi f \in W^{s}(\mathbb{X})$, for any $\phi \in \mathcal{C}^{\infty}(\mathbb{X})$, supported on $B_{3\delta}(y_0)$ and equals one on $B_{3\delta/2}(y_0)$. We can choose ϕ such that $\phi(\mathbb{X}) \subset [0, 1]$. To verify Equation (8), we assume $N \geq \frac{2}{\delta}$ and estimate,

$$\begin{aligned}\sup_{x\in B_\delta(y_0)} |f(x)-\sigma_N(f,x)| &\le \sup_{x\in B_\delta(y_0)} |f(x)-\sigma_N(\phi f,x)| + \sup_{x\in B_\delta(y_0)} |\sigma_N(\phi f - f)| \\ &\le \sup_{x\in B_\delta(y_0)} |\phi(x)f(x)-\sigma_N(\phi f,x)| + \sup_{x\in B_\delta(y_0)} |\sigma_N(\phi f - f, x)| \\ &\lesssim N^{-s} + \sup_{x\in B_\delta(y_0)} |\sigma_N(\phi f - f, x)|,\end{aligned}$$

where we have used global approximation results from Mhaskar (2010), applied to the globally smooth function ϕf. In the remainder of this section, we shall estimate $\sup_{x\in B_\delta(y_0)} |\sigma_N(\phi f - f, x)| N^{-s}$.

To do so, we use the localization property of Φ_N, i.e., for $S > \max(1, \alpha + s)$,

$$|\Phi_N(x,y)| \frac{N^\alpha}{\max\left(1, (N\rho(x,y))\right)^S}$$

(Filbir and Mhaskar, 2010; Maggioni and Mhaskar, 2008; Mhaskar; 2011). This kernel localization yields, for $x \in B_\delta(y_0)$,

$$|\sigma_N(\phi f - f, x)| \lesssim \sum_{j=0}^{n} \omega_j (1-\phi(x_j)) f(x_j) \frac{N^\alpha}{\max\left(1, (N\rho(x,x_j))\right)^S}.$$

We split the sum into two parts, $\mathcal{I}_1 = \{j : x_j \notin B_{3\delta/2}(y_0)\}$ and the remaining indices, on which the summands vanish. Thus, we obtain

$$|\sigma_N(\phi f - f, x)| \lesssim \sum_{j\in\mathcal{I}_1} \omega_j (1-\phi(x_j)) f(x_j) \frac{N^\alpha}{\max\left(1, (N\rho(x,x_j))\right)^S}.$$

Since $N \ge \frac{2}{\delta}$ and $\rho(x, x_j) \ge \delta/2$ on the range that we consider, we obtain

$$|\sigma_N(\phi f - f, x)| \lesssim \sum_{j\in\mathcal{I}_1} \omega_j (1-\phi(x_j)) f(x_j) \frac{N^{\alpha - S}}{\rho(x,x_j)^S}.$$

The ball $B_{\delta/2}(x)$ is contained in $B_{3\delta/2}(y_0)$ so that $\mathcal{I}_2 = \{j : x_j \notin B_{\delta/2}(x)\}$ yields

$$\begin{aligned}|\sigma_N(\phi f - f, x)| &\lesssim \sum_{j\in\mathcal{I}_2} \omega_j (1-\phi(x_j)) f(x_j) \frac{N^{\alpha - S}}{\rho(x,x_j)^S} \\ &\lesssim \sup_{j\in J_2} (f(x_j)) N^{\alpha - S} \sum_{j\in\mathcal{I}_2} \frac{\omega_j}{\rho(x,x_j)^S} \\ &\lesssim N^{\alpha - S} \sum_{j\in\mathcal{I}_2} \frac{\omega_j}{\rho(x,x_j)^S},\end{aligned}$$

where we have used that f is bounded. To finalize the proof, we can apply the crude estimate $\sum_{j\in\mathcal{I}_2} \frac{\omega_j}{\rho(x,x_j)^S} \le \left(\frac{2}{\delta}\right)^S$, which follows from $\sum_{j=1}^{n} \omega_j = 1$. The relation $\alpha - S \le -s$ implies the desired inequality $\sup_{x\in B_\delta(y_0)} |\sigma_N(\phi f - f, x)| \lesssim N^{-s}$, where the constant may depend on f, y_0, δ, α, and s.

Appendix D. Computing eigenfunctions

The graph Laplacian, as an operator on smooth functions, is

$$(\mathcal{L}_M^\varepsilon f)(x) = \frac{1}{M}\left(\sum_{j=1}^{M} k_\varepsilon(x, y_j) f(y_j) - f(x) \sum_{j=1}^{M} k_\varepsilon(x, y_j)\right)$$

and, when $\{y_i\}_{i=1}^M$ is uniformly distributed, $\mathcal{L}_M^\varepsilon$ converges almost surely toward $\Delta_{\mathbb{X}}$ as M tends to infinity and ε to zero (Lafon, 2004; Nadler et al., 2006; Singer, 2006). Let $\lambda_{M,i}^\varepsilon$ and $g_{M,i}^\varepsilon$ be the ith eigenvalue and eigenfunction of $(2\pi\varepsilon)^{1-\alpha/2}\mathcal{L}_M^\varepsilon$, respectively, where α is again the dimension of the underlying manifold $\mathbb{X}$. If λ_i and φ_i are the ith eigenvalue and eigenfunction of $\Delta_{\mathbb{X}}$, then according to Belkin and Niyogi (2006), there exists a sequence $\varepsilon_M \to 0$ such that, in probability,

$$\lim_{M\to\infty} |\lambda^{\varepsilon_M}_{M,i} - \lambda_i| = 0, \qquad \lim_{M\to\infty} \|g^{\varepsilon_M}_{M,i} - \varphi_i\|_{L_2} = 0.$$

In order to compute $g^{\varepsilon}_{M,i}$, we need to relate the spectrum of $\mathcal{L}^{\varepsilon}_{M}$ with the spectrum of the matrix $\mathcal{L}^{\varepsilon}_{M}$ introduced in Section 4. It was shown in von Luxburg et al. (2008) that, for fixed ε and M, the "interesting" eigenvalues and eigenvectors of $\mathcal{L}^{\varepsilon}_{M}$ are in a one-to-one relationship with the "interesting" eigenvalues and eigenfunctions of $\mathcal{L}^{\varepsilon}_{M}$, respectively. In fact, if g is any eigenfunction of $\mathcal{L}^{\varepsilon}_{M}$ with eigenvalue λ, then the sampling $v = (g(y_1), \dots, g(y_M))^\top$ is an eigenvector of $\mathcal{L}^{\varepsilon}_{M}$ with eigenvalue $M\lambda$. If λ is in the discrete spectrum of $\mathcal{L}^{\varepsilon}_{M}$, then g is of the form

$$g(x) = \frac{\sum_{i=1}^{M} k_\varepsilon(x, y_i) v_i}{\sum_{j=1}^{M} k_\varepsilon(x, y_j) - M\lambda}. \tag{9}$$

Conversely, if v is an eigenvector of $\mathcal{L}^{\varepsilon}_{M}$ with eigenvalue $M\lambda$ such that λ is not in the essential spectrum of $\mathcal{L}^{\varepsilon}_{M}$, then g defined by Equation (9) is an eigenfunction of $\mathcal{L}^{\varepsilon}_{M}$ with eigenvalue λ. Note that then g interpolates v, i.e., $(g(y_1), \dots, g(y_M))^\top = v$. Thus, we diagonalize $\mathcal{L}^{\varepsilon}_{M}$ for large M and small ε, and interpolate via Equation (9), which yields an approximation of the eigenfunctions of $\Delta_{\mathbb{X}}$.

Note that the relation between eigenvectors of $\mathcal{L}^{\varepsilon}_{M}$ and $\mathcal{L}^{\varepsilon}_{M}f$ described above are independent on the distribution of $\{y_i\}_{i=1}^{M}$. This is useful if $\{y_i\}_{i=1}^{M}$ are not uniformly distributed but distributed according to the density p. In this case, $\mathcal{L}^{\varepsilon}_{M}f$ approximates the Fokker-Planck operator $\Delta + \frac{\Delta p}{p}$ (Coifman et al., 2005b; Nadler et al., 2006).

Appendix E. Polynomials needed for the example on the sphere

E.1. Gegenbauer polynomials. The Gegenbauer polynomials

$$P_k^{(s)}(x) = \sum_{\ell=0}^{\lfloor k/2 \rfloor} (-1)^\ell \frac{\Gamma(k-\ell+s)}{\Gamma(s)\ell!(k-2\ell)!} (2x)^{k-2\ell},$$

are orthogonal polynomials on the interval $[-1, 1]$ with respect to the weight function $(1 - x^2)^{s-1/2}$, where Γ is the usual Gamma function.

E.2. Orthonormal basis for $\mathcal{H}^d_k$. In the following, we shall present an explicit basis $\{\varphi_{k,i}\}_{i=1}^{m_k}$ for $\mathcal{H}^d_k$ such that we can solve $\sum_{j=1}^{n} \omega_j \varphi_{k,i}(x_j) = \delta_{k,0}$, for all $i = 1, \dots, m_k$ and $k = 0, \dots, R$, so that $N = \sum_{k=0}^{R} m_k$.

For $m, s \in \mathbb{N}$, and $d > 1$, let

$$G^s_m(x_{(d)}) := \sum_{i=0}^{\lfloor m/2 \rfloor} (-1)^i \frac{|x_{(d-1)}|^{2i} x_d^{m-2i}/(m-2i)!}{(2,2)_i (d-1+2s, 2)_i},$$

where $(a, b)_i = a(a+b)\cdots(a+(i-1)b)$ with the convention $(a, b)_0 = 1$ and $x_{(d)} = (x_1, \dots, x_d)$. The collection of polynomials

$$G_\nu(x_{(d)}) := G^{\nu_2}_{\nu_1-\nu_2}(x_{(d)}) \cdots G^{\nu_d}_{\nu_{d-1}-\nu_d}(x_{(d-2)}),$$

for $\nu \in \mathrm{N}^d$, $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_d = 0, 1$, forms an orthonormal basis for $\mathcal{H}^d_{\nu_1}$ (Karachik, 1998). This basis in hand, we can find the weights $\{\omega_j\}_{j=1}^{n}$ that are needed to define $\sigma_N(f, x)$. For quadratures on S^2, see Filbir and Themistoclakis (2008), and for other bases of $\mathcal{H}^d_k$, see, for instance, Dunkl and Xu (2001). Note that the polynomial expression of $G_\nu(x_{(d)})$ can be computed using computer algebra software. Evaluating these polynomials at our data points lead to round-off errors that can be controlled applying standard three-term relations of orthogonal polynomials.

ACKNOWLEDGMENTS

M.E. was supported by the NIH/DFG Research Career Transition Awards Program (EH 405/1-1/575910). The research of F.F. was partially funded by Deutsche Forschungsgemeinschaft grant FI 883/3-1. H.N.M. was supported, in part, by grant DMS-0908037 from the National Science Foundation and grant W911NF-09-1-0465 from the U.S. Army Research Office.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Antoine, J.P., and Vandergheynst, P. 1995. Wavelets on the *n*-sphere and other manifolds. *J. Math. Phys.* 7, 1013–1104.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural. Comput.* 15, 1373–1396.

Belkin, M., and Niyogi, P. 2006. Convergence of laplacian eigenmaps, 129–136, *In* Schölkopf, B., Platt, J.C., and Hoffman, T., ed. *NIPS*, MIT Press, Cambridge, MA.

Bird, A.C., Bressler, N.M., Bressler, S.B., et al. (1995). An international classification and grading system for age-related maculopathy and age-related macular degeneration. The international ARM epidemiological study group. *Surv. Ophthalmol.* 39, 367–374.

Chew, E.Y., Lindblad, A.S., Clemons, T., et al. (2009). Summary results and recommendations from the age-related eye disease study. *Arch. Ophthalmol.* 127, 1678–1679.

Chui, C.K., and Donoho, D.L. (eds.) (2006). Special issue: Diffusion maps and wavelets. *Appl. Comput. Harmon. Anal.* 21.

Chui, C.K., and Wang, J.Z. (2010). Randomized anisotropic transform for nonlinear dimensionality reduction. *International J. on Geomathematic.* 1, 25–50.

Coifman, R.R., Lafon, S., Lee, A.B., et al. (2005b). Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part I: Diffusion maps. *Proc. Nat. Acad. Sci.* 102, 7426–7431.

Coifman, R.R., Lafon, S., Lee, A.B., et al. (2005a). Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part II: Multiscale methods. *Proc. Nat. Acad. Sci.* 102, 7432–7438.

Coleman, H.R., Chan, C., Ferris, F.L., et al. (2008). Age-related macular degeneration. *Lancet.* 372, 1835–1845.

Costa, J.A., and Hero, A.O. (2004). Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing.* 52, 2210–2221.

Dahlke, S., Dahmen, W., Weinreich, I., et al. (1995). Multiresolution analysis and wavelets on S^2 and S^3. *Numer. Funct. Anal. Optim* 16, 19–41.

Dahlke, S., Steidl, G., and Teschke, G. (2004). Coorbit spaces and banach frames on homogeneous spaces with applications to the sphere. *Adv. Comput. Math.*, 21, 147–180.

Detrano, R., Janosi, A, Steinbrunn, W., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology.* 64, 304–310.

Dobrosotskaya, J., Ehler, M., King, E.J., et al. (2010). Sparse representation and variational methods in retinal image processing. *Intern. Fed. for Medical & Biological Engineering*, Springer Proceedings Series. 26th Southern Biomedical Engineering Conference, 361–364.

Dobrosotskaya, J., Ehler, M., King, E.J., et al. (2011). Modeling of the rhodopsin bleaching with variational analysis of retinal images. *Proceedings of SPIE* 7962, 79624N.

Dunkl, C.F., and Xu, Y. (2001). *Orthogonal Polynomials of Several Variables*, Number 81. Encyclopedia of Mathematics and its Applications. Cambridge Univ. Press, New York.

Ehler, M., Dobrosotskaya, J., King, E.J., et al. (2011a). Modeling photo-bleaching kinetics to map local variations in rod rhodopsin density. *SPIE Medical Imaging, Computer-Aided Diagnosis*, 79633R.

Ehler, M., Kainerstorfer, J., Cunningham, D., et al. (2011b). Extended correction model for optical imaging. *IEEE International Conference on Computational Advances in Bio and Medical Sciences.* 93–98.

Ehler, M., Majumdar, Z., King, E.J., et al. (2010). High-resolution autofluorescence imaging for mapping molecular processes within the human retina. *Intern. Fed. for Medical & Biological Engineering*, Springer Proceedings Series. 26th Southern Biomedical Engineering Conference, 344–347.

Filbir, F., and Mhaskar, H.N. (2010). A quadrature formula for diffusion polynomials corresponding to a generalized heat kernel. *J. Fourier Anal. Appl.* 16, 629–657.

Filbir, F., and Mhaskar, H.N. (2011). Marcinkiewicz–Zygmund measures on manifolds. *Journal of Complexity*, 27, 568–596.

Filbir, F. and Themistoclakis, W. (2008). Polynomial approximation on the sphere using scattered data. *Math. Nachr.* 281, 650–668.

Gavish, M., Nadler, B., and Coifman, R.R. (2010). Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. *ICML* 367–374.

Gia, Q.T.L., and Mhaskar, H.N. (2006). Polynomial operators and local approximation of solutions of eudo-differential operators on the sphere. *Numerische Mathematik.* 103, 299–322.

Gia, Q.T.L., and Mhaskar, H.N. (2008a). Localized linear polynomial operators and quadrature formulas on the sphere. *SIAM J. Numer. Anal.*, 47(1):440–466.

Gia, Q.T.L., and Mhaskar, H.N. (2008b). Quadrature formulas and localized linear polynomial operators on the sphere. *SIAM J. Numer. Anal.* 47, 440–466.

Jones, P.W., Maggioni, M., and Schul, R. (2010). Universal local parametrizations via heat kernels and eigenfunctions of the laplacian. *Ann. Acad. Sci. Fenn. Math.* 35, 131–174.

Kainerstorfer, J., Amyot, F., Ehler, M., et al. (2010a). Direct curvature correction for non-contact imaging modalities – applied to multi-spectral imaging. *J. Biomed. Opt.*, 15, 046013.

Kainerstorfer, J., Ehler, M., Amyot, F., et al. (2010b). Principal component model of multi spectral data for near real-time skin chromophore mapping. *J. Biomed. Opt.,* 15, 046007.

Kainerstorfer, J., Riley, J.D., Ehler, M., et al. (2011). Quantitative principal component model for skin chromophore mapping using multi spectral images and spatial priors. *Biomedical Optics Express.* 2, 1040–1058.

Karachik, V.V. (1998). On one set of orthogonal harmonic polynomials. *Proc. Amer. Math. Soc.* 126, 3513–3519.

Krishnadev, N., Meleth, A.D., and Chew, E.Y. (2010). Nutritional supplements for age-related macular degeneration. *Curr. Opin. Ophthalmol.* 21, 184–9.

Lafon, S. (2004). *Diffusion maps and geometric harmonics* [Ph.D. thesis]. Yale University, New Haven, CT.

Maggioni, M., and Mhaskar, H.N. (2008). Diffusion polynomial frames on metric measure spaces. *Appl. Comput. Harmon. Anal.* 24, 329–353.

Meyers, S.M., Ostrovsky, M.A., and Bonner, R.F. (2004). A model of spectral filtering to reduce photochemical damage in age-related macular degeneration. *Trans. Am. Ophthalmol. Soc.* 102, 83–93, discussion 93–5.

Mhaskar, H.N. (2010). Eignets for function approximation. *Appl. Comput. Harmon. Anal.* 29, 63–87.

Mhaskar, H.N. (2011). A generalized diffusion frame for parsimonious representation of functions on data defined manifolds. *Neural Networks.* 24, 345–359.

Nadler, B. and Galun, M. (2007). Fundamental limitations of spectral clustering. *Neural Information Processing systems.* 19, 1017–1024.

Nadler, B., Lafon, S., Coifman, R.R., et al. (2006). Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators. *In* Weiss, Y., Schölkopf, B., and Platt, J., ed. *Adv. Neural Inform. Process. Syst.*, Volume 18. MIT Press, Cambridge, MA.

Pesenson, I.Z., and Pesenson, M.Z. (2010). Sampling, filtering and sparse approximations on combinatorial graphs. *J. Fourier Anal. Appl.* 16, 921–942.

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science.* 290, 2323–2326.

Saito, N. (2008). Data analysis and representation on a general domain using eigenfunctions of laplacian. *Appl. Comput. Harmon. Anal.* 25, 68–97.

Sikora, A. (2004). Riesz transform, gaussian bounds and the method of wave equation. *Math. Z.* 247, 643–662.

Singer, A. (2006). From graph to manifold Laplacian: The convergence rate. *Appl. Comput. Harmon. Anal.* 21, 128–134.

Starck, J.L., Moudden, Q., Abrial, P., et al. (2006). Wavelets, ridgelets and curvelets on the sphere. *Astronomy & Astrophysics.* 446, 1191–U62.

Stein, E., and Weiss, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, Princeton, N.J.

Tenenbaum, J.B., de Silva, V., and Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science.* 290, 2319–23.

van Leeuwen, R., Klaver, C.C., Vingerling, J.R., et al. (2003). The risk and natural course of age-related maculopathy. *Arch. Ophthalmol.* 121, 519–526.

von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of spectral clustering. *Ann. Stat.* 36, 555–586.

Wolberg, W.H., and Mangasarian, O.L. (1990). Multisurface method of pattern separation applied to breast cytology diagnosis. *Proceedings of the National Academy of Sciences.* 87, 9193–9196.

Wu, Q., Guinney, J., Maggioni, M., et al. (2010). Learning gradients: predictive models that infer geometry and dependence. *Journal of Machine Learning Research.* 11, 2175–2198.

Address correspondence to:
Dr. Martin Ehler
Helmholtz Zentrum München
Institute of Biomathematics and Biometry
Ingolstädter Landstrasse 1
D-85764 Neuherberg
Germany

E-mail: martin.ehler@helmholtz-muenchen.de